ISSN 1745-8587

School of Economics, Mathematics and Statistics

BWPEF 0910

Real-time Inflation Forecast Densities from Ensemble Phillips Curves

Anthony Garratt
Birkbeck, University of London

James Mitchell
NIESR

Shaun P. Vahey
Melbourne Business School

Elizabeth C. Wakerly
Melbourne Business School

September 2009

Birkbeck Working Papers in Economics & Finance

Real-time Inflation Forecast Densities from Ensemble Phillips Curves*

Anthony Garratt (Birkbeck College)
James Mitchell (NIESR)
Shaun P. Vahey (Melbourne Business School)
Elizabeth C. Wakerly (Melbourne Business School)

September 7, 2009

Abstract

A popular macroeconomic forecasting strategy takes combinations across many models to hedge against model instabilities of unknown timing; see (among others) Stock and Watson (2004) and Clark and McCracken (2009). In this paper, we examine the effectiveness of recursive-weight and equal-weight combination strategies for density forecasting using a time-varying Phillips curve relationship between inflation and the output gap. The densities reflect the uncertainty across a large number of models using many statistical measures of the output gap, allowing for a single structural break of unknown timing. We use real-time data for the US, Australia, New Zealand and Norway. Our main finding is that the recursive-weight strategy performs well across the real-time data sets, consistently giving well-calibrated forecast densities. The equal-weight strategy generates poorly-calibrated forecast densities for the US and Australian samples. There is little difference between the two strategies for our New Zealand and Norwegian data. We also find that the ensemble modeling approach performs more consistently with real-time data than with revised data in all four countries.

Keywords: Model uncertainty; Density combination; Ensemble forecasting; VAR models
JEL codes: C32; C53; E37

*Financial support from the ESRC (grant No. RES-062-23-1753) and the ARC (grant No. LP0991098) is gratefully acknowledged. Contact: Anthony Garratt, Birkbeck, University of London, Malet Street, Bloomsbury, London, WC1E 7HX, U.K. Tel: +44 (0) 207 631 6410. Fax: +44 (0) 631 6416. E-Mail: a.garratt@bbk.ac.uk. We benefitted greatly from comments by Dean Croushore, Domenico Giannone, James Hamilton, Simon van Norden and David Papell. We are also indebted to conference and seminar participants at the Reserve Bank of Australia and the CIRANO Data Revisions Workshop October 2008.

1 Introduction

In the presence of unknown structural changes, a number of studies have found that forecast combination using time-varying recursive weights, based on historical forecast performance, is an ineffective strategy for improving point forecasts. Stock and Watson (2004) and Clark and McCracken (2009), among others, have found that an equal-weight strategy is more effective in terms of root mean squared forecast error.

Increasingly, forecasters and policymakers in economics are interested in forecast densities rather than point forecasts. Jore, Mitchell and Vahey (2009) [JMV] demonstrate that a recursive weight density combination strategy, using the logarithmic scores of the component forecast densities, produces well-calibrated ensemble densities.

This paper examines the effectiveness of recursive and equal-weight combination strategies for forecast densities using a Phillips curve relationship between inflation and the output gap. We consider convex combinations of probability forecasts known as the "linear opinion pool" (see Timmermann (2006, p177)). In addition to allowing for a large number of candidate detrending methods, we consider a model space which allows for a single structural break of unknown timing. We evaluate the recursive and equal-weight strategies by examining the probability integral transforms (*pits*) of the combined densities for inflation in real time. We consider real-time data for the US, Australia, New Zealand and Norway.

For the recursive-weight strategy, we construct forecast density combinations based on the logarithmic score for inflation forecast densities produced from the component models to generate model weights; see, for example, Garratt, Mitchell and Vahey (2009) [GMV].

We compare and contrast the inflation forecasts from equal-weight ensembles and the recursive-weight variants. The recursive-weight strategy performs well across the real-time data sets, consistently giving well-calibrated forecast densities. The equal-weight strategy performs less consistently, generating poorly-calibrated forecast densities for the US and Australian samples in particular. There is little difference between the two strategies for our New Zealand and Norwegian data. We also find that the ensemble modeling approach performs more consistently with real-time data than with revised data in all four countries.

The remainder of this paper is structured as follows. In Section 2, we outline the component models. In Section 3, we describe our methods for forecast density combination and evaluation. In Section 4, we apply our methodology to US, Australian, New Zealand and Norwegian data and presents the results. In the final section we conclude and discuss the scope for future research.

2 Component models

Following Orphanides and van Norden (2005) and GMV, we consider Phillips curve forecasting models of the form:

$$\pi_{t+h} = \alpha_1^j + \sum_{p=1}^{P} \beta_{1,p}^j \pi_{t-p+1} + \sum_{p=1}^{P} \gamma_{1,p}^j y_{t-p+1}^j + \varepsilon_{1,t+h}^j, \tag{1}$$

where inflation is defined as $\pi_t = \ln(pr_t) - \ln(pr_{t-1})$, where pr is the price level, and the various output gap measures are denoted y_t^j, where $j = 1, \ldots, J$; P denotes the maximum number of lags in inflation and the output gap measures respectively, and h is the forecast horizon.[1] Notice that there is model uncertainty over the output gap measure but also the appropriate values of P (treated as fixed here). We therefore will have N different models $i = 1, \ldots, N$, defined over different values of J, P and, as introduced below, the number of break dates used when considering structural change, each with their associated forecasts of inflation.

We augment this specification with the corresponding output gap equation to create a bivariate VAR system. The output gap equation takes the form:

$$y_{t+h}^j = \alpha_2^j + \sum_{p=1}^{P} \beta_{2,p}^j \pi_{t-p+1} + \sum_{p=1}^{P} \gamma_{2,p}^j y_{t-p+1}^j + \varepsilon_{2,t+h}^j. \tag{2}$$

Note that for simplicity, we have assumed that the lag structure is identical in the two equation system.[2]

To facilitate comparison with GMV, we consider their base set of seven "flexible time trends", derived from univariate filters ($J = 7$). We define the output gap as the difference between observed output and unobserved potential (or trend) output. Let q_t denote the (logarithm of) actual output in period t reported in a given vintage of data dated $t + 1$, and μ_t^j be its trend using definition j where $j = 1, 2, \ldots, J$. Then the output gap, y_t^j, is defined as the difference between actual output and its j^{th} trend measure. We assume the following trend-cycle decomposition:

$$q_t = \mu_t^j + y_t^j.$$

The seven methods of univariate trend extraction are: quadratic [Q], Hodrick-Prescott [HP], a forecast-augmented HP [HPF], Christiano and Fitzgerald [CF], Baxter-King [BK], Beveridge-Nelson [BN], and Unobserved Components [UC]. GMV describe the specifications of each detrending approach.

In our applications the total number of models considered will be determined by the number of variants of P used in equations (1) and (2), as well the number of breaks considered. We allow for lags 1 to 4 for both terms in each equation (*i.e.* $P = 1, 2, 3, 4$), so that in total we consider 28 (7×4) components in the model space before considering structural break variants.

A number of VAR studies have noted the scope for parameter change to improve forecasting performance. Following GMV, we expand the model space to allow for a single structural break of unknown timing in a pragmatic and computationally convenient manner. For each VAR specification described above, we consider every candidate break date, assuming a coincident break in the conditional mean and variance. With the computational burden in mind, we assume the break dates to be identical across equations for

[1] We set $h = 1$ in our applications that follow to simplify the presentation of the results from our many models.

[2] Larger systems pose no conceptual problems but add to the computational burden.

each VAR specification, where the break date is restricted to occur before the start of the evaluation period in which component densities are combined.[3]

3 Methods for ensemble combinations and evaluations

The inflation targeting regimes adopted worldwide, following the innovative steps taken by the Reserve Bank of New Zealand in 1988, focused the attention of many central banks on forecasting inflation. More recently, several central banks (including the Bank of England, Norges Bank and Sveriges Riksbank) have moved to publish forecast densities for key macroeconomic variables. With these developments in mind, we construct ensemble forecast densities for inflation based on the (out of sample) forecast performance of our many component models (described in the previous section).

We construct the predictive densities for the component models using forecast density combination methods. Earlier papers, by JMV and GMV, take this approach to ensemble modeling. In contrast, Stock and Watson (2004) and Clark and McCracken (2009) (among others) study point forecast combinations across a large number of models. Although point forecast combination has a longer tradition in economics (e.g., see Bates and Granger (1969)) than ensemble forecasting, the focus of our study is on providing monetary policymakers with an estimate of the entire probability distribution of the possible future values of the variable of interest—the forecast density.

GMV explore the similarities and differences between the uncertain instabilities literature, typified by macro-econometric work by Stock-Watson and Clark-McCracken, and the ensemble forecasting literature in weather forecasting and climatology. A recent paper by Bache, Mitchell, Ravazzolo and Vahey (2009) describes the embryonic ensemble forecasting literature in macro-econometrics, and provides a characterization. In short, the ensemble methodology combines the forecast densities from a large number of relatively simple component models using time-varying weights to approximate the (likely nonlinear, and non-Gaussian) data generating process. The macro-econometric literature stemming from Stock-Watson (and others) can be distinguished by the emphasis on point forecast combination and simple combination strategies, such as equal weights.

3.1 Ensemble methods

To construct our ensemble forecasts, we aggregate forecasts supplied by "experts". Each expert uses a unique two equation VAR specification to produce a forecast density for inflation.; see (among others) JMV and Timmermann (2006, p177)). Given $i = 1, \ldots, N$

[3]Since we have one break in every feasible observation prior to the start of the evaluation period, which varies by sample, the total number of components varies by country. We report the number of components in each case in the data section.

VAR specifications, the ensemble densities are defined by the convex combination:[4]

$$p(\pi_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h}\; g(\pi_{\tau,h} \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{3}$$

where $g(\pi_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities from model i, $i = 1, \ldots, N$ of inflation π_τ, conditional on the information set I_τ. The publication delay in the production of real-time data ensures that this information set contains lagged variables, here assumed to be dated $\tau - 1$ and earlier. Each individual model is used to produce h-step ahead forecasts via the direct approach; see the discussion by Marcellino, Stock and Watson (2003). Hence, the macro variables used to produce an h-step ahead forecast density for τ are dated $\tau - h - 1$. (In applying this framework to the data from four countries below, we set $h = 1$ for simplicity.) The non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity.[5]

Since each VAR considered produces a forecast density that is multivariate Student-t (see the discussion in Garratt, Koop, Mise and Vahey (2009)), the combined density defined by equation (3) will be a mixture—accommodating skewness and kurtosis. That is, the combination delivers a more flexible distribution than each of the individual densities from which it was derived. As N increases, the combined density becomes more and more flexible, with the potential to approximate non-linear specifications.

We consider two distinct strategies for constructing the weights, $w_{i,\tau,h}$: recursive weights and equal weights. In the former, the weights change with each recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$. In the latter, we restrict the weights to be constant and equal throughout the evaluation.

3.1.1 Recursive weights (RW)

With the RW strategy, we construct the ensemble weights based on the fit of the individual model forecast densities. Like Amisano and Giacomini (2007) and Hall and Mitchell (2007), we use the logarithmic score to measure density fit for each component model through the evaluation period. The logarithmic scoring rule gives a high score to a density forecast that assigns a high probability to the realized value.[6] Specifically, following JMV, the recursive weights for the h-step ahead densities take the form:

$$w_{i,\tau,h} = \frac{\exp\left[\sum_{\underline{\tau}-tr}^{\tau-1-h} \ln g(\pi_{\tau,h} \mid I_{i,\tau})\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-tr}^{\tau-1-h} \ln g(\pi_{\tau,h} \mid I_{i,\tau})\right]}, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau} \tag{4}$$

[4] The linear opinion pool is often justified by considering an expert combination problem. Wallis (2005) proposes the linear opinion pool as a tool to aggregate forecast densities from survey participants. Hall and Mitchell (2007) combine two inflation density forecasts from two institutions.

[5] The restriction that each weight is positive could be relaxed; for discussion see Genest and Zidek (1986).

[6] The logarithmic score of the i-th density forecast, $\ln g(\pi_{\tau,h} \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn $\pi_{\tau,h}$.

where the $\underline{\tau} - tr$ to $\underline{\tau}$ comprises the training period used to initialize the weights. Computation of these weights is feasible for a large N ensemble. Given the uncertain instabilities problem, the recursive weights should be expected to vary across τ.

From a Bayesian perspective, density combination based on recursive logarithmic score weights, RW, has many similarities (for $h = 1$) with an approximate predictive likelihood approach (see Raftery and Zheng (2003), and Eklund and Karlsson (2007)). Given our definition of density fit, the model densities are combined with equal (prior) weight on each model—which a Bayesian would term non-informative priors. Given these weights, we construct an aggregate forecast density for inflation (recursively, at each horizon). GMV use those same weights to construct an h-step ahead ensemble predictive for the output gap.

3.1.2 Equal weights (EW)

The EW approach attaches equal (prior) weights to each model with no updating of the weights through the recursive analysis: $w_{i,\tau,h} = w_{i,h} = 1/N$.

3.2 Forecast density evaluations

In constructing the RW forecast densities, we evaluate forecasts using the logarithmic score at each recursion. The many models are repeatedly evaluated using real-time data. A popular evaluation method for forecast densities, following (for example) Dawid (1984) and Diebold *et al* (1998), evaluates relative to the "true" but unobserved density using the probability integral transforms (*pits*) of the realization of the variable with respect to the forecast densities. A density forecast can be considered optimal (regardless of the user's loss function) if the model for the density is correctly conditionally calibrated; i.e., if the *pits* $z_{\tau,h}$, where:

$$z_{\tau,h} = \int_{-\infty}^{\pi_{\tau,h}} p(u)du,$$

are uniform and, for one-step ahead forecasts, independently and identically distributed. In practice, therefore, density evaluation with the *pits* requires application of tests for goodness-of-fit and independence at the end of the evaluation period.[7]

The goodness-of-fit tests employed include the Likelihood Ratio (LR) test proposed by Berkowitz (2001). We use a three degrees-of-freedom variant with a test for independence, where under the alternative $z_{\tau,h}$ follows an AR(1) process. Since the LR test has a maintained assumption of normality, we also consider the Anderson-Darling (AD) test for uniformity, a modification of the Kolmogorov-Smirnov test, intended to give more weight to the tails (and advocated by Noceti, Smith and Hodges (2003)). We also follow Wallis (2003) and employ a Pearson chi-squared test which divides the range of the $z_{\tau,h}$ into eight equiprobable classes and tests whether the resulting histogram is uniform. To

[7] Given the large number of component densities under consideration, we do not allow for parameter uncertainty when evaluating the *pits*. Corradi and Swanson (2006) review *pits* tests computationally feasible for small N.

test independence of the *pits*, we use a Ljung-Box (LB1) test, based on autocorrelation coefficients up to four (with our quarterly data).

4 Applications

We begin our analysis by describing the sample data for each of our four countries. Then we present the results, focusing on the calibration properties of the inflation forecast densities for the two strategies EW and RW.

4.1 Data

In this section we describe the four samples used, for the US, Australia, New Zealand and Norway. Throughout our analysis, we use real-time observations for real output. We note that the availability of real-time data differs across the countries, with the US and Australian data sets covering longer periods in comparison with New Zealand and Norway. As a result, the evaluation periods are considerably shorter in these two cases. We provide more details below.

United States

For the US, we use the same real-time US data set as Clark and McCracken (2009) and GMV. The quarterly real-time data used refer to real GDP and the GDP price deflator. Here we use 83 vintages (data observed at a specific point in time), starting in 1987q1 ending in 2007q3. The data for each vintage, avoiding the period of the Korean War, are for $1954q3, \ldots, \tau - 1$ where $\tau - 1 = 1986q4, \ldots, 2007q2$. Data on output and the price deflator are first released with a one quarter lag.

The raw data for GDP (in practice, GNP for some vintages) are taken from the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists. This is a collection of vintages of National Income and Production Accounts; each vintage reflects the information available around the middle of the respective quarter. Croushore and Stark (2001) provide a description of the database. The US evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1991q4$ and $\overline{\tau} = 2007q3$ (64 observations), as we drop the first 20 quarters which we use as the training period to intilaise weights (i.e. $tr = 20$ in equation (4)), reflecting the large sample size available in the case of the US. To implement density combination through the evaluation period requires an additional assumption about which measurement is to be forecast. Following Clark and McCracken (2009), GMV and others, we use the second estimate as the "final" data to be forecast. For consistency, we report results for the same definition of "final" data for all forecast density combinations and evaluations. See also the discussion in Corradi, Fernandez and Swanson (2007).

To repeat, for all four countries, in each of our VAR applications, we consider lag lengths of one to four ($P = 1, 2, 3, 4$) and have set $J = 7$. We also allow for a single structural break of unknown timing in each VAR component. The break occurs in the conditional mean and the variance for both equations. This pragmatic treatment of

structural breaks implies that we forecast with VARs using a variety of expanding windows for parameter estimation. The break date is restricted to occur before the start of the evaluation period to reduce the computational burden. When considering structural breaks, at each recursion, the sample size varies between the full sample $1, 2, ..., \tau - 1$ and $0.85 \times (\tau - 1), ..., \tau - 1$ for the US, thereby covering a minimum of 15% of the full sample in each recursion. Hence in the case of the US, we consider 376 component models for each measure of the output gap considered. With seven measures of the output gap derived from flexible trends, the predictives combine 2632 component specifications for each observation in the evaluation period.

Australia

Our real-time real output data for Australia were obtained from the Reserve Bank of Australia (described in detail in Reserve Bank of Australia Research Discussion Papers No 2002-05 and 2002-06). There are 63 vintages of quarterly real GDP seasonally-adjusted data starting in 1991q3 and ending in 2007q4. The data for each vintage are for $1959q3, \ldots, \tau - 1$ where $\tau - 1 = 1991q2, \ldots, 2007q3$ where data on output are first released with a one quarter lag. As real-time data for prices is not available for Australia (this is also true of New Zealand and Norway), we use the consumer price index from a single vintage, as a feature of consumer prices is that they typically have very minimal revisions. The consumer price series was downloaded from the IMF's International Financial Statistics data base, dated July 2009. The Australian evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1996q2$ and $\overline{\tau} = 2007q4$ (47 observations). As the sample sizes were comparable to those used for the US, structural breaks were handled in an identical manner i.e. $tr = 20$ and we require a minimum of 15% of the sample, for each recursion, for all regressions. Hence the number of component models for each measure of the output gap is 356, making for a total of 2492 models evaluated in the combination each period.

New Zealand

Our real-time real output data for New Zealand were obtained from the Reserve Bank of New Zealand (described in detail at www.rbnz.govt.nz/research/2482495.html). There are 40 vintages of quarterly real GDP seasonally-adjusted data starting in 1998q1 and ending in 2007q4, where the data for each vintage are for $1987q2, \ldots, \tau - 1$ where $\tau - 1 = 1997q4, \ldots, 2007q3$. Data on output are first released with a one quarter lag. The consumer price series was downloaded from the IMF's International Financial Statistics data base in July 2009. The New Zealand evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1999q1$ and $\overline{\tau} = 2007q4$ (36 observations), as given the shorter sample, we drop just 5 observations as the training period to intialise weights ($tr = 5$). Similar considerations were also applied when dealing with structural breaks, where at each recursion, the sample size varies between the full sample $1, 2, ..., \tau - 1$ and $0.50 \times (\tau - 1), ..., \tau - 1$, hence we have a minimum of 50% of the sample. As a consequence the number of component models for each measure of the output gap is 48, making for a total of 336 models evaluated in the combination each period.

Norway

The real-time real output data for Norway were obtained from Norges Bank.[8] There are 29 vintages of quarterly real GDP seasonally-adjusted data starting in 2001q2 and ending in 2008q2, where the data for each vintage are for $1978q1, \ldots, \tau - 1$ where $\tau - 1 = 2001q1, \ldots, 2008q1$. Data on output are first released with a one quarter lag. The consumer price series was downloaded from the IMF's International Financial Statistics data base in July 2009. The Norwegian evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 2002q2$ and $\overline{\tau} = 2008q2$ (25 observations). For Norway we also drop 5 observations as the training period (i.e $tr = 5$) and the restrict the sample size when considering structural break to be a minimum of 50%. Hence the number of component models for each measure of the output gap is 208, making for a total of 1456 models evaluated in the combination each period.

4.2 Results

In this section, we present our results on the calibration properties of the forecast densities resulting from our ensemble methodology, for the EW and RW strategies. We begin with the US results (which we treat separately on the grounds that the real-time data are of exceptional quality), and then turn to the remaining three countries.

4.2.1 US

Table 1 reports the *pits* tests $p-$values, together with the AD test statistic (which has a 95% critical value of 2.5). The figures in bold denote that the forecast density is correctly calibrated for a 95% confidence interval on the basis of that individual test; that is, when we cannot reject at a 95% confidence level the null hypothesis that the densities are correctly calibrated. There are four rows to the table. The first two refer to the RW strategy, with real time data (RW-RT), and final-vintage data (RW-FV), respectively. The third and fourth rows give corresponding results for the EW strategy, with real-time data (EW-RT), and final-vintage data (EW-FV), respectively.

Looking at the real time data results, we see that the RW strategy gives well-calibrated densities on the basis of all seven tests, row 1. But the EW strategy fails two of the seven tests in real time, row 3.

Turning to the revised final-vintage data, we see that for both the EW and RW strategies, the performance is somewhat weaker. The RW strategy passes four of the seven tests, and EW passes three.

[8] They can be obtained from Norges Bank on request.

Table 1: US Ensembles

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW-RT	**0.19**	**0.38**	**0.29**	**0.33**	**1.02**	**0.16**	**0.24**
RW-FV	0.01	**0.12**	**0.27**	0.01	**2.08**	0.01	**0.29**
EW-RT	0.02	**0.52**	**0.10**	0.03	**1.96**	**0.19**	**0.11**
EW-FV	0.00	**0.07**	**0.09**	0.00	3.01	0.04	**0.06**

Notes: LR2 is the p-value for the Likelihood Ratio test of zero mean and unit variance of the inverse normal cumulative distribution function transformed *pits*, with a maintained assumption of normality for the transformed *pits*; LR_{upper} is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent upper tail; LR_{lower} is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent lower tail; LR3 supplements LR2 with a test for zero first order autocorrelation; AD is the Anderson-Darling test statistic for uniformity of the *pits* which assuming independence of the *pits* has an associated 95 percent asymptotic critical value of 2.5. χ^2 is the p-value for the Pearson chi-squared test of uniformity of the *pits* histogram in eight equiprobable classes. LB is the p-value from a Ljung-Box test for independence of the *pits*. The Log Score is the average log score over the evaluation period.Notes: LR2 is the p-value for the Likelihood Ratio test of zero mean and unit variance of the inverse normal cumulative distribution function transformed *pits*, with a maintained assumption of normality for the transformed *pits*; LR_{upper} is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent upper tail; LR_{lower} is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent lower tail; LR3 supplements LR2 with a test for zero first order autocorrelation; AD is the Anderson-Darling test statistic for uniformity of the *pits* which assuming independence of the *pits* has an associated 95 percent asymptotic critical value of 2.5. χ^2 is the p-value for the Pearson chi-squared test of uniformity of the *pits* histogram in eight equiprobable classes. LB is the p-value from a Ljung-Box test for independence of the *pits*. The Log Score is the average log score over the evaluation period.

4.2.2 Australia, New Zealand and Norway

Tables 2, 3 and 4 present the results for Australia, New Zealand and Norway, respectively. The Australian results, in Table 2, suggest that like the US case, the RW strategy produces well-calibrated real-time densities (row 1). Although we note that the RW strategy does fail two of the seven tests. In contrast, the EW strategy indicates calibration failure for five of the seven tests (row 3). As with the US results, we see weaker calibration for both strategies with final-vintage data. For example, with the EW strategy, row 4, the null of correct calibration is not rejected on the basis of just one test with 95 percent confidence.

Turning to the New Zealand and Norway results, Tables 3 and 4, respectively, we see that RW and EW perform similarly in real time. The RW strategy results in only one rejection at a 95 percent confidence interval (row 1 in each table). And the EW strategy betters that slightly with no tests failed (row 3 in each case).

As for the Australian and US data, both strategies fail more tests with final-vintage data. For example, with the EW strategy, five and four tests are failed for New Zealand

and Norway, respectively; see row 4. For the RW strategy, five (New Zealand) and three (Norway) tests are failed, respectively; see row 2.

Table 2: Australian Ensembles

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW-RT	**0.417**	**0.353**	0.013	**0.428**	**2.280**	0.011	**0.970**
RW-FV	**0.159**	**0.061**	0.002	**0.172**	3.198	0.014	**0.164**
EW-RT	0.015	**0.437**	0.013	0.021	4.032	0.008	**0.940**
EW-FV	0.016	0.014	0.010	0.027	4.477	0.010	**0.195**

Notes: see notes to Table 1

Table 3: New Zealand Ensembles

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW - RT	**0.085**	**0.556**	**0.684**	**0.107**	2.523	**0.230**	**0.479**
RW - FV	0.000	0.022	**0.204**	0.000	2.907	0.004	**0.696**
EW - RT	**0.071**	**0.361**	**0.936**	**0.074**	**2.265**	**0.333**	**0.508**
EW - FV	0.000	0.022	**0.149**	0.000	3.136	0.003	**0.474**

Notes: see notes to Table 1.

Table 4: Norwegian Ensembles

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW - RT	**0.185**	**0.537**	0.030	**0.280**	**1.624**	**0.689**	**0.385**
RW - FV	0.019	**0.064**	0.033	0.028	**1.530**	**0.611**	**0.249**
EW - RT	**0.154**	**0.142**	**0.254**	**0.157**	**1.392**	**0.611**	**0.132**
EW - FV	0.028	0.035	0.022	0.041	**1.205**	**0.766**	**0.341**

Notes: see notes to Table 1.

4.2.3 Interpretation

Overall, there are two substantive findings. First, we see that for the relatively short New Zealand and Norwegian samples, there is little to separate the EW and RW strategies, with both strategies giving real-time forecast densities that cannot reject the null of no calibration failure. In contrast, for the longer US and Australian real-time samples, the EW strategy fails a number of *pits* tests. The RW strategy seems more robust on the these longer real-time samples.

The second findings is that the density forecasting performance is less satisfactory for the ensembles with final-vintage data. Data revisions contaminate the Phillips curve relationship in all four countries considered.

5 Conclusions

In this paper, we have examined the effectiveness of recursive-weight and equal-weight strategies for combining forecast densities using a Phillips curve relationship between inflation and the output gap. Using data for the US, Australia, New Zealand and Norway, we find that the recursive weight strategy performs consistently well. In the two cases with longer samples of real time data—the US and Australia—the equal-weight strategy results in forecast densities that exhibit calibration failure. This result reverses the perceived wisdom that simple averages are more reliable—a result found in a number of well-known studies of point forecasting accuracy. In future work, we intend to investigate the calibration properties of recursive-weight ensembles with dynamic stochastic general equilibrium components.

References

[1] Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Likelihood Ratio Tests", *Journal of Business and Economic Statistics*, 25, 2, 177-190.

[2] Bache, I.W., J. Mitchell, F. Ravazzolo and S.P. Vahey (2009), "Macro Modelling with Many Models", NIESR Discussion Paper No. 337.

[3] Bates, J.M., and C.W. Granger (1969), "The Combination of Forecasts", *Operations Research Quarterly*, 20, 451-468.

[4] Berkowitz, J. (2001), "Testing Density Forecasts, with Applications to Risk Management", *Journal of Business and Economic Statistics*, 19, 465-474.

[5] Clark, T.E. and M. W. McCracken (2009) "Averaging Forecasts from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, 24, 7, forthcoming.

[6] Corradi, V. and N.R. Swanson (2006), "Predictive Density Evaluation", G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting,* North-Holland, 197-284.

[7] Corradi, V., A. Fernandez and N.R. Swanson (2007), "Information in the Revision Process of Real-time Data", Downloadable at http://econweb.rutgers.edu/nswanson/papers/revision_final_06_02_2009.pdf. Forthcoming *Journal of Business and Economic Statistics.*

[8] Croushore, D. and T. Stark. (2001), "A Real-time Data Set for Macroeconomists", *Journal of Econometrics,* 105, 111-130.

[9] Dawid, A.P. (1984), "Statistical Theory: The Prequential Approach", *Journal of the Royal Statistical Society B*, 147, 278-290.

[10] Diebold, F.X., Gunther, T.A. and A.S. Tay (1998) "Evaluating Density Forecasts; with applications to financial risk management", *International Economic Review,* 39, 863-83.

[11] Eklund, J. and S. Karlsson (2007), "Forecast Combination and Model Averaging Using Predictive Measures", *Econometric Reviews*, 26(2-4), 329-363.

[12] Garratt, A, Koop, G., Mise, E. and S.P. Vahey (2009), "Real-Time Prediction with UK Monetary Aggregates in the Presence of Model Uncertainty", downloadable at http://www.ems.bbk.ac.uk/faculty/garratt, forthcoming *Journal of Business and Economic Statistics.*

[13] Garratt, A, Mitchell, J. and S.P. Vahey (2009), "Measuring Output Gap Uncertainty", Birkbeck College Mimeo, downloadable at http://www.ems.bbk.ac.uk/faculty/garratt.

[14] Genest, C. and J.V. Zidek (1986) "Combining Probability Distributions: A Critique and an Annotated Bibliography", *Statistical Science,* 1, 114-148.

[15] Hall, S.G. and J. Mitchell (2007), "Combining Dnsity Forecasts", *International Journal of Forecasting,* 23, 1-13.

[16] Jore, A. S., J. Mitchell and S.P. Vahey (2009), "Combining Forecast Densities from VARs with Uncertain Instabilities", NIESR Discussion Paper No. 303, forthcoming *Journal of Applied Econometrics.*

[17] Marcellino, M., J. Stock and M.W. Watson (2003), "A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Series h-steps ahead", *Journal of Econometrics*, 135, 499–526.

[18] Nocetti, P., Smith, J. and S. Hodges (2003), "An Evaluation of Tests of Distributional Forecasts *Journal of Forecasting*, 22, 447-455.

[19] Orphanides, A. and S. van Norden (2005) "The Reliability of Inflation Forecasts Based on Output-Gap Estimates in Real Time", *Journal of Money Credit and Banking,* 37, 3, 583-601.

[20] Raftery, A.E. and Y. Zheng (2003), "Long-Run Performance of Bayesian Model Averaging *Journal of the American Statistical Association*, 98, 931-938.

[21] Stock, J.H. and M.W. Watson (2004) "Combination Forecasts of Output Growth in a Seven-Country Data Set", *Journal of Forecasting*, 23, 405-430.

[22] Timmermann, A. (2006), "Forecast Combination", G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting,* North-Holland, 197-284.

[23] Wallis, K.F. (2003), "Chi-squared Tests of Interval and Density Forecasts, and the Bank of England's Fan Charts", *International Journal of Forecasting*, 19, 165-175.

[24] Wallis, K.F. (2005) "Combining Density and Interval Forecasts: a Modest Proposal", *Oxford Bulletin of Economics and Statistics*, 67, 983-994.